Proceedings of "The 6th SEAMS-UGM Conference 2011"
Topic, pp. xx–xx.

SEQUENCE ANALYSIS OF DNA H1N1 VIRUS USING SUPER PAIRWISE ALIGNMENT

ALFI YUSROTIS ZAKIYYAH, M.ISA IRAWAN, MAYA SHOVITRI

Abstract. Sequence analysis is one of methods to assign function, structure evolution and features from sequence and one of kind this methods is a Super Pairwise Alignment. This methods was assigned homologous sequences DNA H1N1 Virus.

Keywords and Phrases : DNA, Homologous, Super Pairwise Aligment

1. Introduction

DNA sequencing methodology was developed in the late 1970 and has become one of the most widely uses technique in moleculer biology. The importance of this technique is underlined by the volume research funds now being invested in development of outomated sequencers and sequence analysis system. Sequence analysis in moleculer biology includes a very wide range of relevan topics like construction of map, translation, protein analysis, similarty search, alignment with a similar sequence and submission and retrievel. The most task sequence analysis is alignment with similar sequence. An optimal alignment is achieved between two similar sequences (DNA or amino acid) and the percent or similarity calculated [1].

One of methods in sequence alignment is dynamic programming. The widely used alignment, dynamic programming though generating optimal alignment, takes too much due to its high computation complexity $O(N^2)$. A majority of sequence alignment software utilizes dynamic programming, such as global Needleman Wunsch and local alignment use Smith Waterman [3]. Both methods were a classical algorithm in sequence alignment. Based on the result of the research of Shen et all, both methods have disadvantage of which one is the speed of computation. To solve this problem, Shen et all found a new methods that is a Super Pairwise alignment. This methods combines the analysis of the methods combinatorics and probability[2,3].

The management of the new virus can be analyzed by homologous. On the problem of the identification of disease, DNA virus mutates so that it may give rise to new viruses. The case H1N1 is one example of mutation. The H1N1 virus mutates quickly enough. Hemaglutinin virus transmission from aspartic acid (D) to Glynine (G) on the line 222 [4]. This becomes the reference to the writer for further consideration applied method Super pairwise Alignment to analyze sequence DNA H1N1 virus.

1.1 Super Pairwise Alignment

The mathematical methods used to study mutation and alignment fall mainly into three groups there are stochastics analysis, modulus structure and combinatorial graph theory [2]. At first glance, a DNA sequence structure may seem disorderly and unsystematics and nucleotides at each positions (or a group of positions) are not fixed. That is to say that biological sequence analysis is stochastics sequence. statistically, we may find that frequency of observing all molecules or segment changes based o different dataset of biological sequences. Therefore, we may use stochastics model to describes biological sequences

.

1.1.1 Parameter Estimation

The key to solving the uniform alignment of the pairwise sequences is knowing how to estimate the parameters in the mutation mode *T* based on sequence (*A,B*). then *T* is a group of statistical parameter and

$$\widehat{T} = \left\{ \left(\hat{i}_k, \hat{\ell}_k \right), k = 1, 2, \ldots \hat{k}_a \right\}$$

is a set of statistics determined by (*A,B*), and estimate of the parameter set *T*. the vital problem of uniform alignment of pairwise sequences is the estimate of the parameter in *T*. The approach to solving this problem is briefly described below :

a. To estimate the parameters in T alternately, we estimate $(i_k, l_k), k = 1, 2, \ldots, k_a$ one after the other, that is, we estimate $(i_{k'}, l_{k'})$ based on $(i_k, l_k), k = 1, 2, \ldots, k' - 1$.
b. To estimate each (i_k, l_k), we need not have the entire data of sequence (*A,B*), but depends on only part of the data. Therefore, choosing the data to use becomes one of the most important aspect of the statistical decision algorithm
c. The estimate of the parameter set *T* includes an estimate of the parameter k_a

1.1.2 Algorithm

Let (*A,B*) be two fixed sequences. This algorithm based on Shen et all research [2]. Typically, we choose n = 20,50, 100, 150 etc. θ, θ' are selected based on the error rate of mutation and error rate of the indepently random variables. The SPA algorithm is described below :

a. Estimate the first mutation position i_1 in T. i is position of mutation and ℓ is length of mutation. Inialize $i = j = 0$ and calculate $w(A, B; i, j, n)$. If $w(A, B; i, j, n) = w \geq \theta'$ then let $\hat{i}_1 = 0$. Otherwise expand value of i
b. Estimate ℓ based on the estimation $\hat{i}_1$ of the first mutation position in T
c. After obtaining the estimation $(\hat{i}_1, \hat{\ell}_1)$, we continue to estimate $\hat{i}_2$ based on C_1, D_1

d. Estimate $\hat{\ell}_2$ based on the estimation $\hat{i}_1, \hat{\ell}_1, \hat{i}_2$
e. The process will be terminate at some C_{k_0} and D_{k_0} have shifting mutation occurring in (A_{2,k_0}, B_{2,k_0}).

1.2 DNA Virus H1N1

In this research, we get sequences of H1N1 from database The national Center for Biotechnology Information [5]. for preliminary research, we use ClustalW software to alignment several strain DNA H1N1 virus. The output from this software is phylogenetics trees which showing the inferred evolutionary relationships among various biological spesies or other entities based upon similarities and differences in their physical and genetic characteristics. The result of running the program Clustal W to determine the phylogenetics tree as follows :

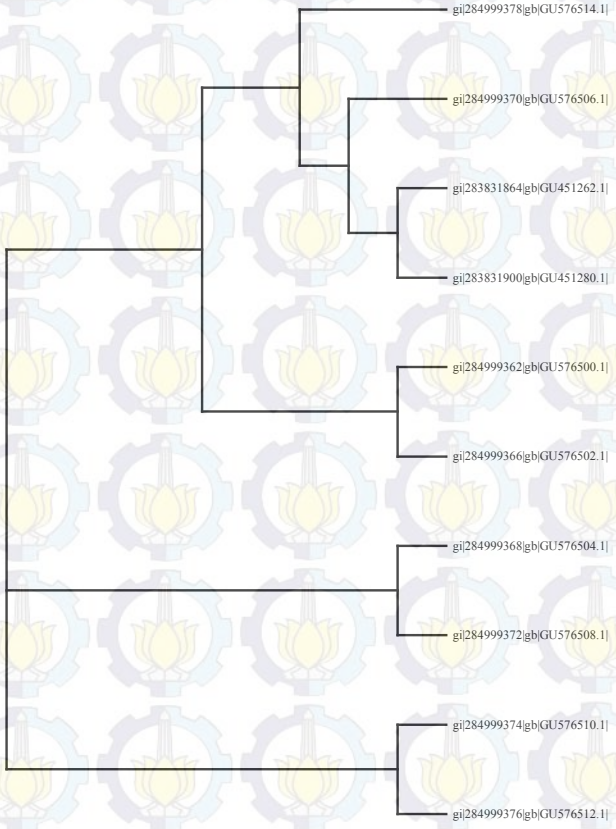

Figure 1. Phylogenetic Tree

2011 Mathematics Subject Classification : Computer Science For Biology and other aturall science

2.SEQUENCE ALIGNMENT DNA H1N1 VIRUS USING SPA

In this section discussion about alignment of several DNA H1N1 virus using Super Pairwise Aligment (SPA). One of pairwise alignment are DNA GU451262 and GU451280 sequences which each length of character 923 bp and 909 respectively. These DNA taken from NCBI [5]. For the first step, take local similarity (n) and for this this problem this study use value of $n_0 = 20$ and $\theta' = 0.6$ and this study obtain the result of local similarity from 20 character of sequences are :

GU451262	G	T	A	G	A	C	A	C	A	G	T	A	C	T	A	G	A	A	A	A
GU451280	T	A	G	A	C	A	C	A	G	T	A	C	T	A	G	A	A	A	A	G

List 2.1

From the list 2.1, There are three pair sequences in the same character from 17nd up to 19nd. Further identify is determined sliding window (w) and compared by $\theta' = 0.6$. If the value of $w \geq \theta'$ and its can be assumed the shifting Mutation in $i = j = 0$ Otherwise, meaning no shifting in [1,n] and it can be continued to estimate $\hat{i}$. in this case, alignment GU451262 dan GU451280, similarity = 3 and $w = 0.85 \geq \theta' = 0.6$. Second step, after it determine local mutation, next step is estimate (ℓ) based on $\hat{i}$ and criterion of $|\ell| \leq \mu$. The result from running program is gived in the list below :

ℓ	-5	-4	-3	-2	-1	0	1	2	3	4	5
w	0.7	0.65	0.75	0.6	0.6	0.85	0	0.85	0.6	0.65	0.7

List 2.2

From the list above can be determined the suitable ℓ value. The suitable ℓ=1 and it is related by $w = 0 \leq \theta = 0.4$. This study assign shifting distance $\ell = 1$ and insert one '-' into the first point of second sequence . so the sequence GU451262 and Sequence GU451280 change into

```
GTAGACACAGTACTAGAAAAGAATGTAACAGTAACACACTCTGTTAACCTTCTAGAAGACAAGCATAACGGGAAACTATGCAAACTAAGAGGGGTAGCCCCATTGCATTTGGGTAAATGTAACATTGCTGGCTGGATCCTGGGAAATCCAGAGTG
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
-TAGACACAGTACTAGAAAAGAATGTAACAGTAACACACTCTGTTAACCTTCTAGAAGACAAGCATAACGGGAAACTATGCAAACTAAGAGGGGTAGCCCCATTGCATTTGGGTAAATGTAACATTGCTGGCTGGATCCTGGGAAATCCAGAGTG
```

In this case, determination one gap in the first position, it doesn't alignment again because fom this step it get optimal alignment. The output result are 14 gap and homolog 99,56%. In the other result alignment using BLAST software, it is obtain 98% homolog with same value of gap. The alignment both sequence, it is related by Clustal W, GU451262 and GU451280 sequence is located in one subdivision.

Comparing result of SPA with BLAST software

Sequence Alignment	BLAST	SPA
- GU451262.1 =923 bp - GU451280.= 909 bp	a. similarity sequence : 906 b. Prosentase homolog :98%	a. similarity sequence :906 b.Prosentase homolog : 99.56%
- GU576500 = 1701 bp - GU576502.= 1701bp	a. Similarity sequence : 1699 b. Prosentase homolog :99%	a. similarity sequence : 1699 b. Prosentase homolog :99,88%
- GU576502= 1701 bp - GU576504.= 1701bp	a. Similarity sequence : 1698 b. Prosentase homolog :98%	a. similarity sequence a : 1698 b. Prosentase homolog :98,5%
- GU576510 = 1701 bp - GU576512.= 1701 bp	a. similarity sequence : 1701 b. Prosentase homolog :100%	a. sequence : 1701 b. Prosentase homolog :100%

List 2.3

In the apply algoritm of Super Pairwise Alignment, some parameter need adjustment as the value of the decision local similarity (n). Inaccuracy when taking value of local similarity (n) influence in to optimization of the sequence alignment. Taking same value of the parameter alignment sequences before, it is used to align GU576500 and GU451280 and it doesn't obtain optimal alignment. These DNA have each character 1701 bp and 909 bp respectively. The alignment both sequences can be obtained the result of local similarity from 20 character of sequences are :

GU576500	A	T	G	A	A	G	G	C	A	A	T	A	C	T	A	G	T	A	G	T
GU451280	T	A	G	A	C	A	C	A	G	T	A	C	T	A	G	A	A	A	A	G

List 2.4

From the list 2.4, there are three pair sequences in the same character at 3nd, 4nd and 18nd position. The result of sliding windows $w = 0.85 \geq \theta' = 0.6$ so it can be conclude the shifting mutation in $i = j = 0$. Next, this study determinate the value of ℓ. The result of ℓ, it same with the value of aligment betwen GU451262 and GU251280 that value of $w = 0 \leq \theta = 0.4$. This study know shifting distance $\ell = 1$ and insert one '-' into the first point of second sequence . By addition with one gap at the first point, aligment both sequences have 231 same characters. Comparing with the alignment by BLAST software there is difference. Using BLAST software, there are 97 gap at the first point and the final result of similarity is 907.

3. CONCLUDING REMARK

Super Pairwise Alignment get optimal alignment. This study found that sequences GU451262 and GU451280 have 99,56% homologous. When both sequences alignt with BLAST software have result about 85%. This study still faced several problem for example how to determinate parameter of local similarity. Determination of local similarity influence the value of optimal alignment.

Acknowledgement. I would like to express my gratitude to Department for Higher Education with scholarship. It is very useful to me to continue my study and research in Institute Teknologi Sepuluh Nopember (ITS) Surabaya.

References

[1] Giffin, Hugh G and Annette M. Griffin., *Computer Analysis of Sequence Data.* Humana Press, Totowa, 1994.

[2] Puzelli, Simona. Marcia Facchinin, Domenico spagnolo, Maria A.De Marco, Laura Calzonetti, Alessandro Zanetti, Roberto Fumaggalli, Maria L.tanzi, Antonio Cassone, Giovannie Rezza, Isabella Donatelli, and The surveillance group for Pandemic A (H1N1) 2009., *Transmission of Hemaglutinin D222G Mutant Strain of Pandemic (H1N1) 2009 Virus*. Vol t6. No,5 May 2010.

[3] Shen, Shi Yi Nankai and Tuszynki., *Theory and Mathematical methods for Bioinformatics.* Springer,New York,2008

[4] Shen, Shi Yi., Jun Yang, Adam Yao, Pei Ing Hwang. *Super Pairwise Alignment (SPA) : An Efficient Approach to Global Alignment For Homologous Sequences. Journal of Computational Biology Volume 9,* Number 3,2002@Mary Ann Liebert inc Pp477-486

[5] Database Sequences DNA H1N1 Virus, National Center for Biotechnology Information (2011) www.ncbi.nlm.nih.gov

ALFI YUSROTIS ZAKIYYAH:Graduate Student of Mathematics Department at Institut Teknologi Sepuluh Nopember.(ITS)
E-mails: yusrotis@gmail.com

M.ISA IRAWAN:Supervisor, Lecturer of Mathematics Department at Institut TeknologiSepuluh Nopember (ITS)
E-mails: mii@its.ac.id

MAYA SHOVITRI: co.Supervisor, Lecturer of Biology Department at Institut Teknologi Sepuluh Nopember (ITS)
E-mails: maya@bio.its.ac.id